Life Expectancy Analysis

Isha Doshi

2022-12-12

**Life expectancy is a statistical measure of the average time someone is expected to live, based on the year of their birth, current age and other demographic factors including their sex. Period life expectancy assumes mortality rates remain constant into the future, while cohort life expectancy uses projected changes in future mortality rates. Period life expectancy (ex) is the average number of additional years a person would live if he or she experienced the age-specific mortality rates of the given area and time period for the rest of their life.**

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyverse)

## ── Attaching packages  
## ───────────────────────────────────────  
## tidyverse 1.3.2 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.5  
## ✔ tibble 3.1.8 ✔ stringr 1.4.1  
## ✔ tidyr 1.2.1 ✔ forcats 0.5.2  
## ✔ readr 2.1.3   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)

**Loaded and cleaned the data–removed all cases with missing life expectancy, year, country name and code.**

data=read.delim("gapminder.csv.bz2")  
ncol(data)

## [1] 25

nrow(data)

## [1] 13055

#renaming time to year  
data=data %>%rename(year=time)  
head(data)

## iso3 name iso2 region sub.region intermediate.region  
## 1 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 2 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 3 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 4 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 5 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 6 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## year totalPopulation fertilityRate lifeExpectancy childMortality  
## 1 1960 54211 4.820 65.662 NA  
## 2 1961 55438 4.655 66.074 NA  
## 3 1962 56225 4.471 66.444 NA  
## 4 1963 56695 4.271 66.787 NA  
## 5 1964 57032 4.059 67.113 NA  
## 6 1965 57360 3.842 67.435 NA  
## youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP\_PC accessElectricity  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## agriculturalLand agricultureTractors cerealProduction fertilizerHa co2  
## 1 NA NA NA NA 11092.67  
## 2 20 NA NA NA 11576.72  
## 3 20 NA NA NA 12713.49  
## 4 20 NA NA NA 12178.11  
## 5 20 NA NA NA 11840.74  
## 6 20 NA NA NA 10623.30  
## greenhouseGases co2\_PC pm2.5\_35 battleDeaths  
## 1 NA 204.6204 NA NA  
## 2 NA 208.8228 NA NA  
## 3 NA 226.1181 NA NA  
## 4 NA 214.8004 NA NA  
## 5 NA 207.6158 NA NA  
## 6 NA 185.2040 NA NA

#checking nulls  
sum(is.na(data$lifeExpectancy))

## [1] 1325

sum(is.na(data$year))

## [1] 36

sum(is.na(data$name))

## [1] 0

sum(is.na(data$iso3))

## [1] 0

sum(is.na(data$iso2))

## [1] 0

#removing nulls and blanks  
data = data[!(is.na(data$lifeExpectancy) | data$lifeExpectancy==""), ]  
data = data[!(is.na(data$year) | data$year==""), ]  
data = data[!(data$name==""), ]  
data = data[!(data$iso3==""), ]  
data = data[!(data$iso2==""), ]  
ncol(data)

## [1] 25

nrow(data)

## [1] 11558

**There are 203 unique countries in our data**

length(unique(data$name))

## [1] 203

**The first and last year with valid life expectancy data**

first = min(data$year)  
firstRow=data[which.min(data$year),]  
firstRow

## iso3 name iso2 region sub.region intermediate.region  
## 1 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## year totalPopulation fertilityRate lifeExpectancy childMortality  
## 1 1960 54211 4.82 65.662 NA  
## youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP\_PC accessElectricity  
## 1 NA NA NA NA NA  
## agriculturalLand agricultureTractors cerealProduction fertilizerHa co2  
## 1 NA NA NA NA 11092.67  
## greenhouseGases co2\_PC pm2.5\_35 battleDeaths  
## 1 NA 204.6204 NA NA

last = max(data$year)  
lastRow=data[which.max(data$year),]  
lastRow

## iso3 name iso2 region sub.region intermediate.region  
## 60 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## year totalPopulation fertilityRate lifeExpectancy childMortality  
## 60 2019 106314 1.901 76.293 NA  
## youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP\_PC accessElectricity  
## 60 NA NA NA NA 100  
## agriculturalLand agricultureTractors cerealProduction fertilizerHa co2  
## 60 NA NA NA NA NA  
## greenhouseGases co2\_PC pm2.5\_35 battleDeaths  
## 60 NA NA NA NA

cat("first year with valid life expectancy", first,"\n")

## first year with valid life expectancy 1960

cat("last year with valid life expectancy ", last,"\n")

## last year with valid life expectancy 2019

**Lowest and highest life expectancy values and the country/year they correspond to**

The lowest life expectancy wass present in Cambodia (1977) The highest life expectancy was present in San Marino (2012)

min=data[which.min(data$lifeExpectancy),]  
min

## iso3 name iso2 region sub.region intermediate.region year  
## 6098 KHM Cambodia KH Asia South-eastern Asia 1977  
## totalPopulation fertilityRate lifeExpectancy childMortality  
## 6098 7196042 5.557 18.907 260.2  
## youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP\_PC  
## 6098 NA NA NA NA  
## accessElectricity agriculturalLand agricultureTractors cerealProduction  
## 6098 NA 25500 1233 1080000  
## fertilizerHa co2 greenhouseGases co2\_PC pm2.5\_35 battleDeaths  
## 6098 NA 73.34 11996.91 0.01019 NA NA

max=data[which.max(data$lifeExpectancy),]  
max

## iso3 name iso2 region sub.region intermediate.region year  
## 10582 SMR San Marino SM Europe Southern Europe 2012  
## totalPopulation fertilityRate lifeExpectancy childMortality  
## 10582 32105 1.26 85.41707 2.4  
## youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP\_PC  
## 10582 NA NA NA 49939.01  
## accessElectricity agriculturalLand agricultureTractors cerealProduction  
## 10582 100 10 NA NA  
## fertilizerHa co2 greenhouseGases co2\_PC pm2.5\_35 battleDeaths  
## 10582 NA NA NA NA NA NA

**The shortest life expectancy corresponds to a genocide in Cambodia which resulted in the death of 1.5 to 2 million people during 1975 to 1979**

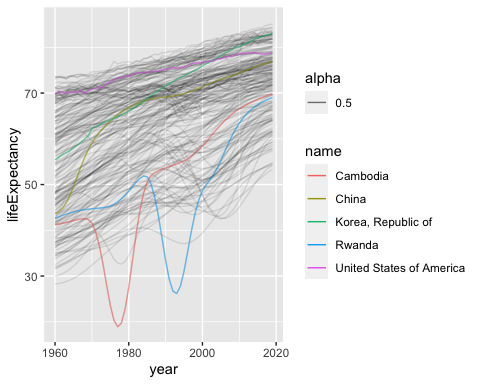
**Plotting the life expectancy over time for all countries.**

**I added Rwanda because there was a genocide in 1994 which resulted in deaths of 800,000 people.**

p <- ggplot(data=data, aes(x=year, y=lifeExpectancy, group=name, fill="gray")) +  
 geom\_line(alpha=0.1)  
data\_subset=data%>%filter(name=="United States of America"|name=="Korea, Republic of"|name=="Cambodia"|name=="China"|name=="Rwanda")  
head(data\_subset)

## iso3 name iso2 region sub.region intermediate.region year totalPopulation  
## 1 CHN China CN Asia Eastern Asia 1960 667070000  
## 2 CHN China CN Asia Eastern Asia 1961 660330000  
## 3 CHN China CN Asia Eastern Asia 1962 665770000  
## 4 CHN China CN Asia Eastern Asia 1963 682335000  
## 5 CHN China CN Asia Eastern Asia 1964 698355000  
## 6 CHN China CN Asia Eastern Asia 1965 715185000  
## fertilityRate lifeExpectancy childMortality youthFemaleLiteracy  
## 1 5.756 43.725 NA NA  
## 2 5.905 44.051 NA NA  
## 3 6.062 44.783 NA NA  
## 4 6.206 45.972 NA NA  
## 5 6.320 47.592 NA NA  
## 6 6.385 49.549 NA NA  
## youthMaleLiteracy adultLiteracy GDP\_PC accessElectricity agriculturalLand  
## 1 NA NA 191.9572 NA NA  
## 2 NA NA 141.0355 NA 3432480  
## 3 NA NA 132.0776 NA 3460010  
## 4 NA NA 142.1449 NA 3488540  
## 5 NA NA 164.1333 NA 3517060  
## 6 NA NA 187.4367 NA 3555090  
## agricultureTractors cerealProduction fertilizerHa co2 greenhouseGases  
## 1 NA NA NA 780726.3 NA  
## 2 52661 109659976 7.04082 552066.8 NA  
## 3 55360 120421293 9.59845 440359.0 NA  
## 4 59657 137456233 12.11821 436695.7 NA  
## 5 66290 152356625 16.32832 436923.0 NA  
## 6 73021 162156281 25.41529 475972.9 NA  
## co2\_PC pm2.5\_35 battleDeaths  
## 1 1.17038 NA NA  
## 2 0.83605 NA NA  
## 3 0.66143 NA NA  
## 4 0.64000 NA NA  
## 5 0.62565 NA NA  
## 6 0.66552 NA NA

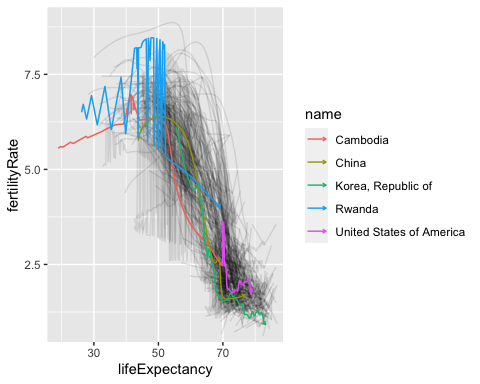
p=p+geom\_line(data=data\_subset, aes(x=year, y=lifeExpectancy,group=name, color=name, alpha=0.5))  
p

 **Life expectancy seems to be increasing over the years, probably due to better better health care and hygiene, healthier lifestyles, diet, and improved medical care. China’s life expectancy improved greatly during the 70s. United States and Korea too has had a better life expectancy over the years. There are dips in Cambodia and Rwanda’s life expectancy due to genocide and tragic killings in the country.**

**Looking at how life expectancy and fertility are related. Made a fertility rate versus life expectancy plot of all countries with selected countries highlighted. Used arrows to mark which way the time goes on the figure.**

plot <- ggplot(data=data, aes(x=lifeExpectancy, y=fertilityRate, group=name, fill="gray")) +  
 geom\_line(alpha=0.1,arrow = arrow())  
plot=plot+geom\_line(data=data\_subset, aes(x=lifeExpectancy, y=fertilityRate, group=name, color=name),arrow = arrow(length=unit(0.10,"cm")))  
plot

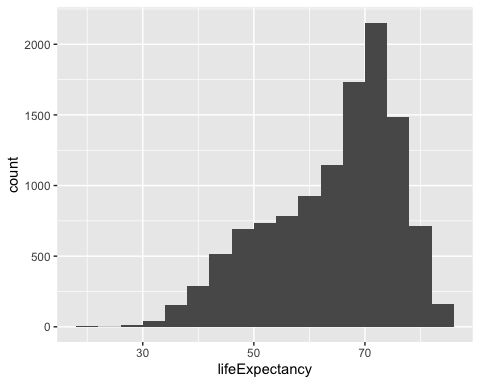
## Warning: Removed 13 row(s) containing missing values (geom\_path).



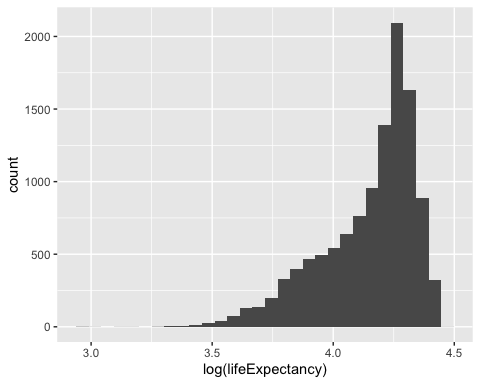
**Fertility rate is decreasing while life expectancy is increasing over time. The reason why fertility rate could be decreasing might be because of women empowerment in education and the workforce, lower child mortality and the increased cost of raising children. The highlighted countries are also following the same trend.**

**Displaying the distribution of life expectancy. It is a little left skewed. We can try log-transformation to see if it distributes the data more normally. Log transformation is making it more left skewed, it would be better to not perform log transformation in this case.**

library(ggplot2)  
ggplot(data, aes(x=lifeExpectancy)) +  
 geom\_histogram(binwidth=4)



library(ggplot2)  
ggplot(data, aes(x=log(lifeExpectancy))) +  
 geom\_histogram(bins=30)

 **Created a model to explain life expectancy with just time, where t is time (year). Used year − 2000 instead of just year for time.Since the data has data points far from each other, scaling technique will help make them closer to each other or in simpler words, scaling will make the data points generalized so that the distance between them will be lower. If the difference between the data points is very high, the model could be unstable, which would result in the model producing poor results. Another reason why this makes more sense is the intercept comes as negative without changing the year, and since life expectancy cannot be negative, it makes sense to scale the data.**

data$mod\_year=data$year-2000  
head(data)

## iso3 name iso2 region sub.region intermediate.region  
## 1 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 2 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 3 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 4 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 5 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## 6 ABW Aruba AW Americas Latin America and the Caribbean Caribbean  
## year totalPopulation fertilityRate lifeExpectancy childMortality  
## 1 1960 54211 4.820 65.662 NA  
## 2 1961 55438 4.655 66.074 NA  
## 3 1962 56225 4.471 66.444 NA  
## 4 1963 56695 4.271 66.787 NA  
## 5 1964 57032 4.059 67.113 NA  
## 6 1965 57360 3.842 67.435 NA  
## youthFemaleLiteracy youthMaleLiteracy adultLiteracy GDP\_PC accessElectricity  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## agriculturalLand agricultureTractors cerealProduction fertilizerHa co2  
## 1 NA NA NA NA 11092.67  
## 2 20 NA NA NA 11576.72  
## 3 20 NA NA NA 12713.49  
## 4 20 NA NA NA 12178.11  
## 5 20 NA NA NA 11840.74  
## 6 20 NA NA NA 10623.30  
## greenhouseGases co2\_PC pm2.5\_35 battleDeaths mod\_year  
## 1 NA 204.6204 NA NA -40  
## 2 NA 208.8228 NA NA -39  
## 3 NA 226.1181 NA NA -38  
## 4 NA 214.8004 NA NA -37  
## 5 NA 207.6158 NA NA -36  
## 6 NA 185.2040 NA NA -35

model<-lm(lifeExpectancy~mod\_year,data=data)  
summary(model)

##   
## Call:  
## lm(formula = lifeExpectancy ~ mod\_year, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.382 -7.605 2.549 8.025 18.524   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.409244 0.109537 615.40 <2e-16 \*\*\*  
## mod\_year 0.309587 0.005457 56.73 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.15 on 11556 degrees of freedom  
## Multiple R-squared: 0.2179, Adjusted R-squared: 0.2178   
## F-statistic: 3219 on 1 and 11556 DF, p-value: < 2.2e-16

**b0 here is 67.40, which is the life expectancy when the year is 0, and b1 is 0.30 which is the coefficient of how year parameter affects the life expectancy.**

**Moving to multiple regression: Estimated the model where I also add the continent (variable region)**

model1<-lm(lifeExpectancy~mod\_year+region,data=data)  
summary(model1)

##   
## Call:  
## lm(formula = lifeExpectancy ~ mod\_year + region, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.161 -4.072 0.549 4.032 20.104   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 55.890766 0.124278 449.72 <2e-16 \*\*\*  
## mod\_year 0.305604 0.003585 85.23 <2e-16 \*\*\*  
## regionAmericas 15.931152 0.183175 86.97 <2e-16 \*\*\*  
## regionAsia 12.206582 0.170420 71.63 <2e-16 \*\*\*  
## regionEurope 20.890772 0.181252 115.26 <2e-16 \*\*\*  
## regionOceania 13.630385 0.265561 51.33 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.666 on 11552 degrees of freedom  
## Multiple R-squared: 0.6625, Adjusted R-squared: 0.6623   
## F-statistic: 4535 on 5 and 11552 DF, p-value: < 2.2e-16

**The region dummies are Americas, Asia, Europe and Oceania. The reference category is Africas. The p value for time trend is <2e-16. The time trend is statistically significant as the probability is less than 0.05. This model performs better than the previous one, since the r square value is higher here.**

**Added two additional variables to the model: log of GDP per capita, and fertility rate. Estimated this model.This model performs better as the adjusted R square is 0.8485, which is higher the the previous two models.**

model2<-lm(lifeExpectancy~mod\_year+region+fertilityRate+log(GDP\_PC),data=data)  
summary(model2)

##   
## Call:  
## lm(formula = lifeExpectancy ~ mod\_year + region + fertilityRate +   
## log(GDP\_PC), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.3227 -2.4592 0.2857 2.7112 12.2179   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.028329 0.507479 98.58 <2e-16 \*\*\*  
## mod\_year 0.138053 0.003539 39.01 <2e-16 \*\*\*  
## regionAmericas 5.939633 0.160004 37.12 <2e-16 \*\*\*  
## regionAsia 5.750250 0.150353 38.24 <2e-16 \*\*\*  
## regionEurope 5.292292 0.207486 25.51 <2e-16 \*\*\*  
## regionOceania 5.665935 0.224681 25.22 <2e-16 \*\*\*  
## fertilityRate -2.250470 0.046215 -48.70 <2e-16 \*\*\*  
## log(GDP\_PC) 2.496868 0.046916 53.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.33 on 8930 degrees of freedom  
## (2620 observations deleted due to missingness)  
## Multiple R-squared: 0.8486, Adjusted R-squared: 0.8485   
## F-statistic: 7150 on 7 and 8930 DF, p-value: < 2.2e-16

**All betas are statistically significant. Fertility rate intercept is now negative. The region dummy values have changed a bit. Europe was the leading region before, but now Americas is leading the pack in terms of the value.Additional variables made the ranking of the continents look different as each additional variable brings new beta which alters how the parameters are interacting with the dependent variable.**

**Based on the most recent model, Americas has the highest life expectancy followed by Asia then Oceania then Europe. We come to this conclusion from the beta values.**